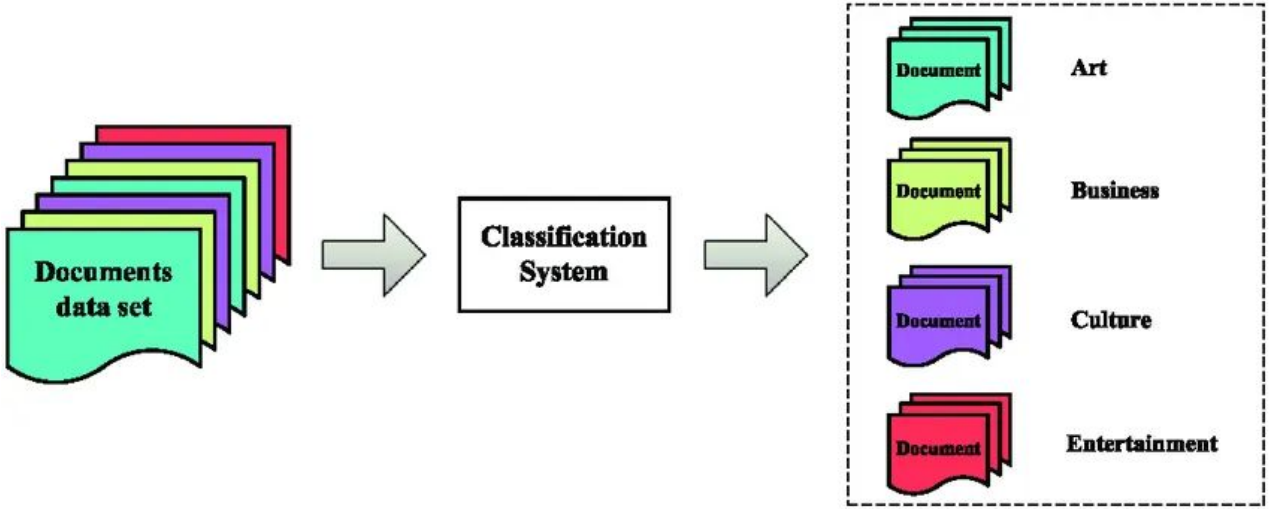
Hebrew Articles Category Classification

Baraa Saleh - 206466732
Marah Aboud - 208865907

Business Model

Our end goal is to create a classification model, that is able to classify articles into their respective category by leveraging data analytics and machine learning techniques

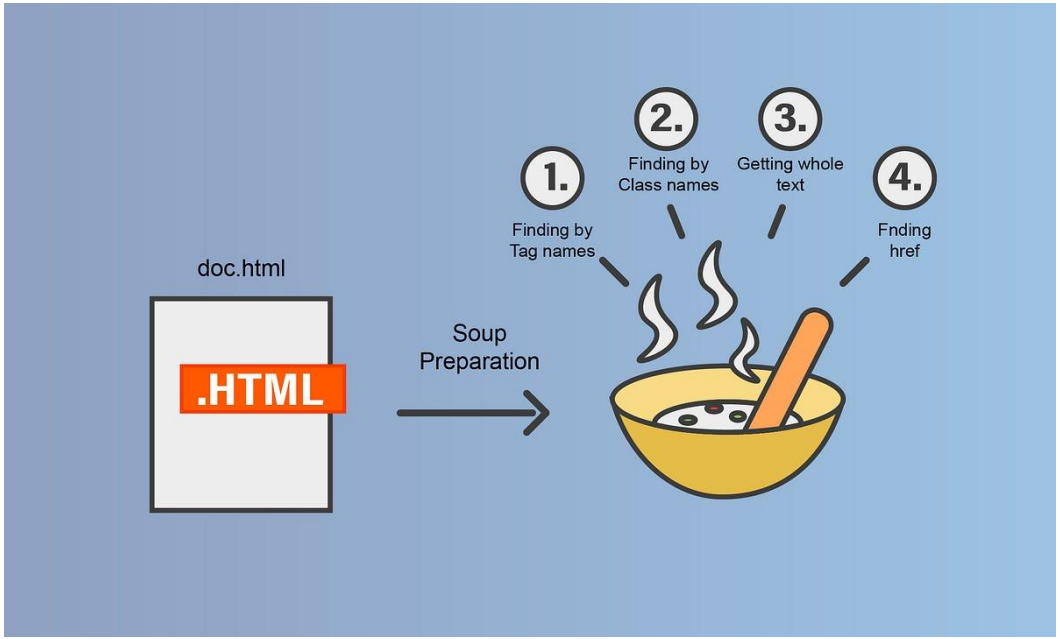


Roadmap

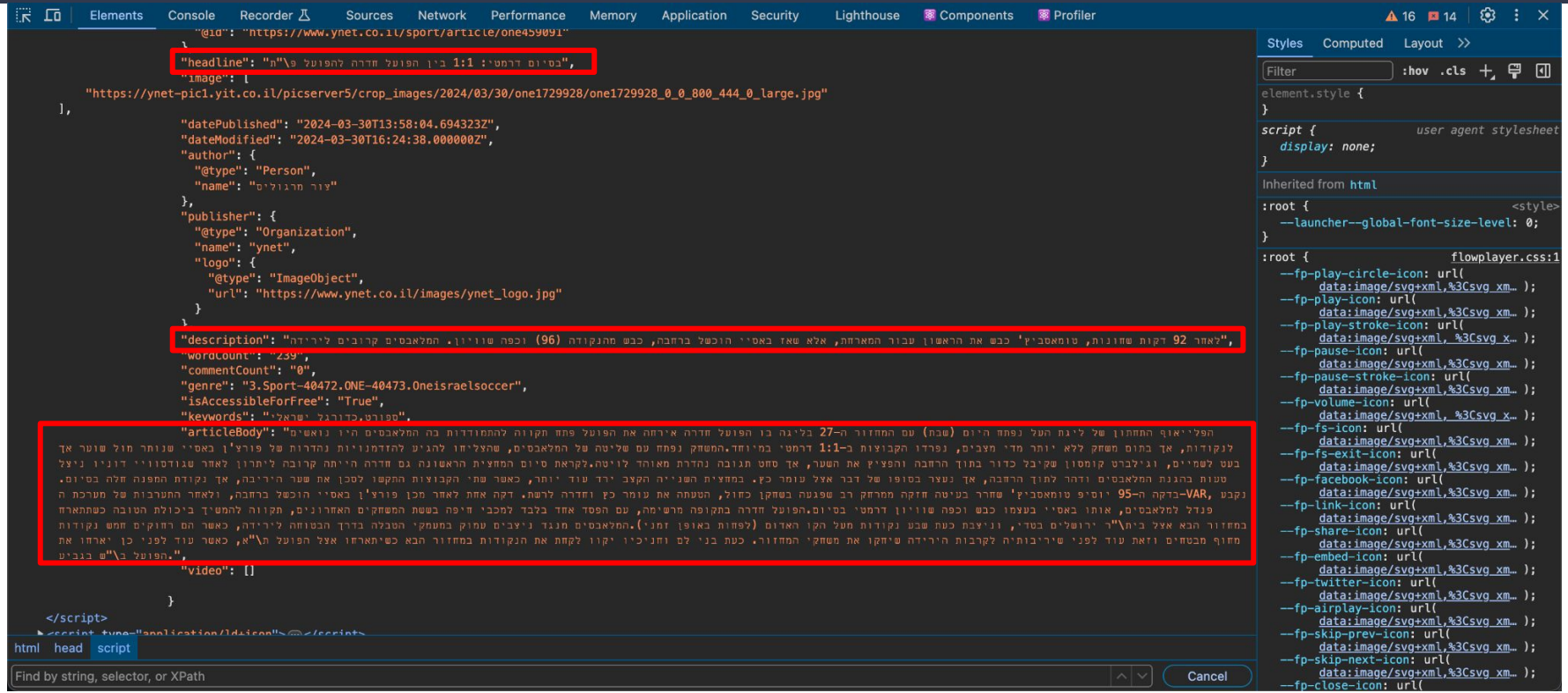
- Data scraping
- Preprocessing
- Classification model
- Backend: Flask
- Frontend: React

Data Scraping: Ynet + BeautifulSoup

We used Ynet which is a popular website containing articles in hebrew and parse the HTML using bs4 (BeautifulSoup) package.



Data Scraping: Headline + Description + Body



Data Scraping: Save the data

Save the article text and its respective category into a Pandas DataFrame and then save the DataFrame into a csv.

	A	B	C	D	E	F	G	H	I	J	K
1	text	sport	entertainment	economy	health	car	food	vacation	dating	parents	environment-science
2	ההסתבכות של מנפורד הצרה כוכב הפועל תא נכשל בבדיקו	1	0	0	0	0	0	0	0	0	0
3	החלוץ והמטרה עם רוזומה מתעתע ובלי דין דוד לצדו פרנצד	1	0	0	0	0	0	0	0	0	0
4	האומץ של קין את כל הבעיות של מכבי תא אפשר היה לראו	1	0	0	0	0	0	0	0	0	0
5	שפרעם הפסידה 21 לכפר קאסם גיא דהן הבריק דלויה שרט	1	0	0	0	0	0	0	0	0	0
6	מתן רודיטי ובקי פודרנר זכו באליפות ישראל במשחה ל500	1	0	0	0	0	0	0	0	0	0
7	בדיחת האולסטאר באופן ברור כפי שנכתב בESPN אדם ס	1	0	0	0	0	0	0	0	0	0
8	קבלו אותה גורבנקו נחתה עם מדליית הכסף ההיסטורית מ	1	0	0	0	0	0	0	0	0	0
9	הביתרים החדשים הממן ברק יצחקי המנהל המקצועי אלמ	1	0	0	0	0	0	0	0	0	0
10	התעוררו עם שחר ערן זהבי סופסל ומכבי תא מצאה סקורר	1	0	0	0	0	0	0	0	0	0
11	שיר כאב איך המלחמה תשפיע על התרבות שלנו הזוועות ה	0	1	0	0	0	0	0	0	0	0
12	זו לא תערוכה על השואה אלא על חוויית הצפייה בזיכרון הש	0	1	0	0	0	0	0	0	0	0
13	מדוע כדאי לכם לחלוץ נעליים בכניסה לבית או שלא רגע לפ	0	1	0	0	0	0	0	0	0	0
14	מקצצים בכבשת הרש של התרבות הספריות הציבוריות אינ	0	1	0	0	0	0	0	0	0	0
15	The Greater Wings של גולי ברן אלבום שמכיר בצער אך	0	1	0	0	0	0	0	0	0	0
16	כשהטמפרטורות יורדות זו המוזיקה הישראלית שתרצו להת	0	1	0	0	0	0	0	0	0	0
17	אמני ישראל לא יוכלו לעבוד בהתנדבות לנצח כולם מתקשים	0	1	0	0	0	0	0	0	0	0
18	סקוט פילגרים הולך על זה ייתן לכם הפוגה מרעשי הרקע ה	0	1	0	0	0	0	0	0	0	0
19	אושר המתוזה של השר מיקי זוהר לתמיכה בנופי התרבות ב	0	1	0	0	0	0	0	0	0	0
20	חמל האמנים עשינו 10 אלף הופעות צריכים עזרה כלכלית כ	0	1	0	0	0	0	0	0	0	0
21	החדש של רולינג סטונס יעניק לכם הפוגה נחוצה המלצות ס	0	1	0	0	0	0	0	0	0	0
22	שנת הקאמבק של טבע נאבקנו שנים כדי לא להפסיד עכשיו	0	0	1	0	0	0	0	0	0	0
23	אני שורף כסף ששמתי בצד לימים קשים אבל לא חשבתי ש	0	0	1	0	0	0	0	0	0	0
24	המילואימניקים שהופקרו בפעם הבאה אנשים יחשבו פעמיי	0	0	1	0	0	0	0	0	0	0
25	מהיום אין מנכל או ממלא מקום לביטוח הלאומי ולא צפוי מנ	0	0	1	0	0	0	0	0	0	0
26	טורקיה יצאה מהמשחק כמה עולה השתלת שיער בישראל ע	0	0	1	0	0	0	0	0	0	0
27	זה העסק שלי 11 שנה אבל בפעם הראשונה שעשיתי גרפיט	0	0	1	0	0	0	0	0	0	0
28	על רקע התקפות החותים CMA CGM מפסיקה שוב לשוט	0	0	1	0	0	0	0	0	0	0
29	הייתי שף במסעדות מישלן באה גברת לחדר אוכל ואמרה ה	0	0	1	0	0	0	0	0	0	0
30	עזיבה נוספת באוצר סגן הממונה על התקציבים הודיע על ס	0	0	1	0	0	0	0	0	0	0
31	גרושה עם שלושה ילדים ועסק חדש אני קצת פנטזיונרית וק	0	0	1	0	0	0	0	0	0	0
32	בזבוז זמן וכסף הפריקטור האזרחי מסיים את תפקידו מבל	0	0	1	0	0	0	0	0	0	0
33	מחלוקת סביב מכרז להפעלת מערך בריאות התלמיד בבתיר	0	0	0	1	0	0	0	0	0	0

Preprocessing: NLP

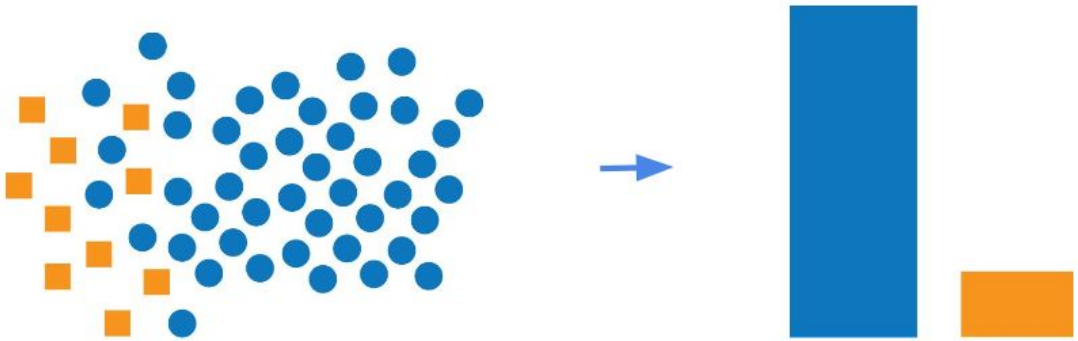
Analyze and clean the data using NLP techniques, such as removing the stop words from the texts which do not add to the understanding of the text.

For example: גם ,או ,אם ,הוא ,כל ,עם ,זה ,על ,אני ,של ,לא ,את, ...

Preprocessing: Imbalanced Data

While checking for imbalances in the data, we noticed that 8 categories had ~900 articles and for 2 categories there were < 50 articles so we removed these 2 categories from the data due to lack of support.

Final data size ~7000 articles



Preprocessing: Encode the categories

Replace all the categories with an encoding to describe the category this text belongs to.

For example:

0 - Sport

1 - Entertainment

2 - Economy

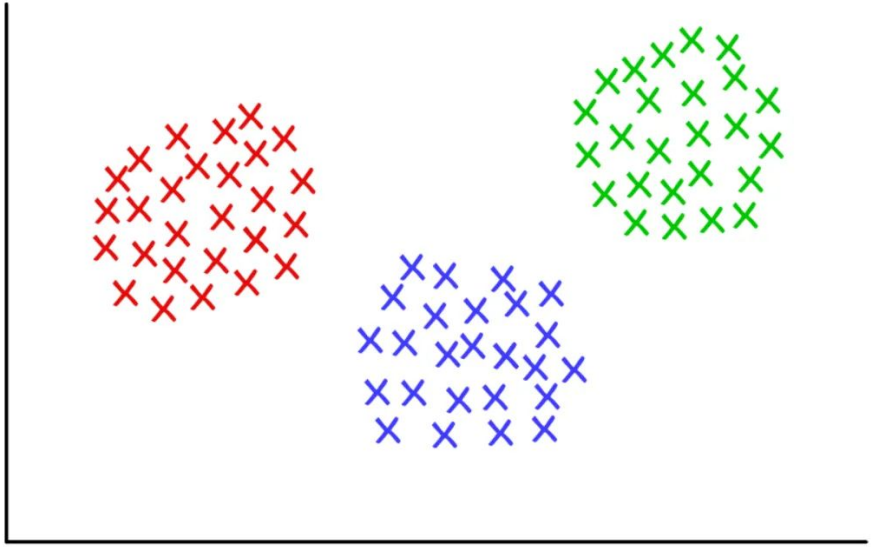
etc ...

	A	B
1	text	category_encoded
2	ההסתבכות של מנפורד הצרה כוכב הפועל תא נכשל בבדיקו	0
3	החלוץ והמטרה עם רזומה מתעתע ובלי דין דוד לצדו פרנצדי	0
4	האומץ של קין את כל הבעיות של מכבי תא אפשר היה לראו	0
5	שפרעם הפסידה 21 לכפר קאסם גיא דהן הבריק דלויה שרע	0
6	מתן רודיטי ובקי פוזדנר זכו באליפות ישראל במשחה ל1500	0
7	בדיחת האולסטאר באופן ברור כפי שנכתב בESPN אדם סי	0
8	קבלו אותה גורבנקו נחתה עם מדליית הכסף ההיסטורית מק	0
9	הביתרים החדשים המאמן ברק יצחקי המנהל המקצועי אלמ	0
10	התעוררו עם שחר ערן זהבי סופסל ומכבי תא מצאה סקורר	0
11	שיר כאב איך המלחמה תשפיע על התרבות שלנו הזוועות ה	1
12	זו לא תערוכה על השואה אלא על חווית הצפייה בזיכרון השו	1
13	מדוע כדאי לכם לחלוץ נעליים בכניסה לבית או שלא רגע לפ	1
14	מקצצים בכבשת הרש של התרבות הספריות הציבוריות אינ	1
15	The Greater Wings של גולי ברן אלבום שמכיר בצער אך	1
16	כשהטמפרטורות יורדות זו המוזיקה הישראלית שתרצו להת	1
17	אמני ישראל לא יוכלו לעבוד בהתנדבות לנצח כולם מתקשים	1
18	סקוט פילגרים הולך על זה ייתן לכם הפוגה מרעשי הרקע ה	1
19	אושר המתווה של השר מיקי זוהר לתמיכה בגופי התרבות ב	1
20	חמל האמנים עשינו 10 אלף הופעות צריכים עזרה כלכלית כ	1
21	החדש של רולינג סטונס יעניק לכם הפוגה נחוצה המלצות ס	1
22	שנת הקאמבק של טבע נאבקנו שנים כדי לא להפסיד עכשיו	2
23	אני שורף כסף ששמתי בצד לימים קשים אבל לא חשבתי שז	2
24	המילואימניקים שהופקרו בפעם הבאה אנשים יחשבו פעמיי	2
25	מהיום אין מנכל או ממלא מקום לביטוח הלאומי ולא צפוי מר	2
26	טורקיה יצאה מהמשחק כמה עולה השתלת שיער בישראל ש	2
27	זה העסק שלי 11 שנה אבל בפעם הראשונה שעשיתי גרפיט	2
28	על רקע התקפות החותים CMA CGM מפסיקה שוב לשוט	2
29	הייתי שף במסעדות מישלן באה גברת לחדר אוכל ואמרה ה	2
30	עזיבה נוספת באוצר סגן הממונה על התקציבים הודיע על סי	2
31	גרושה עם שלושה ילדים ועסק חדש אני קצת פנטזיונרית וקנ	2
32	בזבוז זמן וכסף הפרויקטור האזרחי מסיים את תפקידו מבל	2
33	מחלוקת סביב מכרז להפעלת מערך בריאות התלמיד בבתיר	3

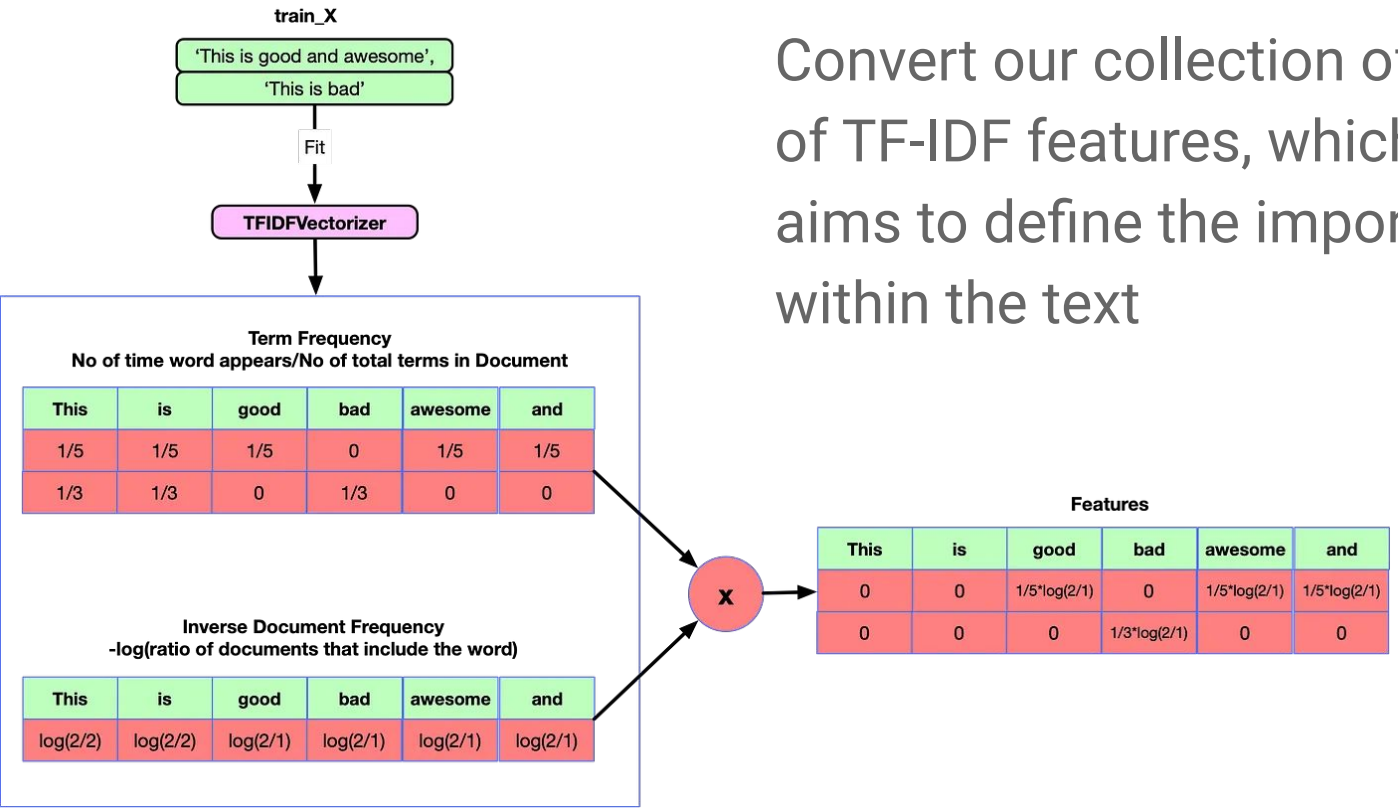
Multiclass Classification

Classification is the problem of assigning observations to one or more categories.

Multiclass classification involves more than 2 categories.



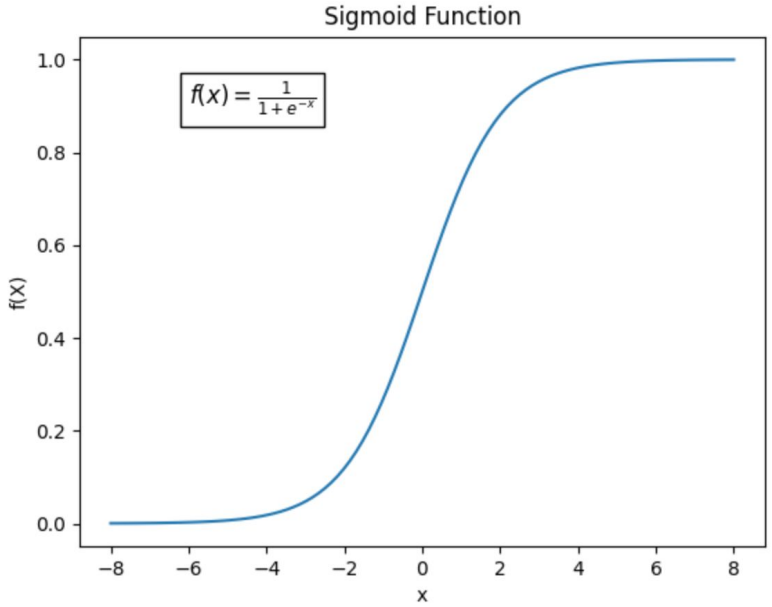
TfidfVectorizer

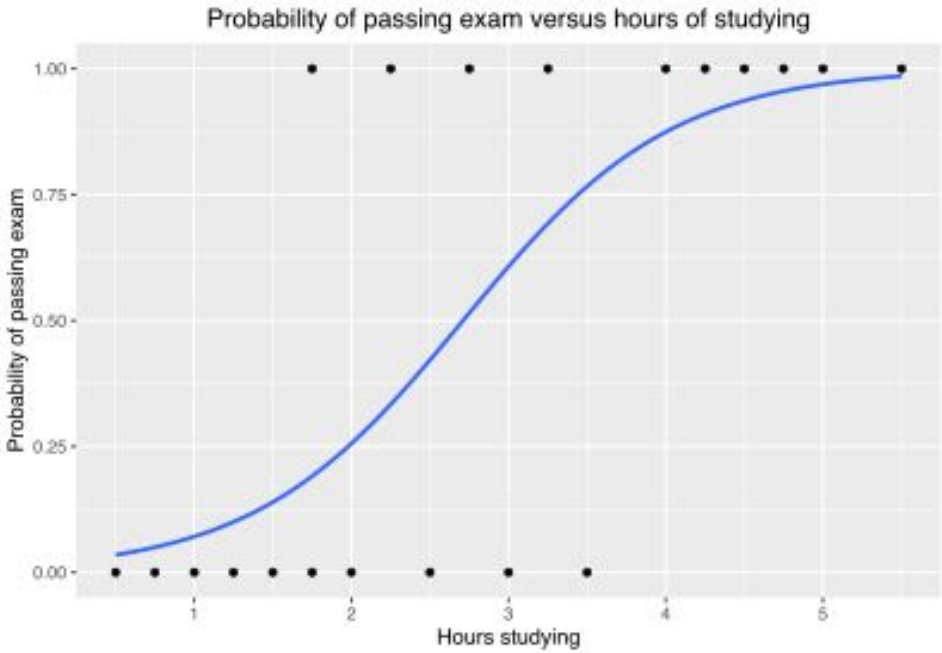


Convert our collection of texts to a matrix of TF-IDF features, which is a formula that aims to define the importance of a word within the text

Classification Model: LogisticRegression

LogisticRegression is a model based on sigmoid function, and here is an example of 2 classes:





Classification Model: LogisticRegression

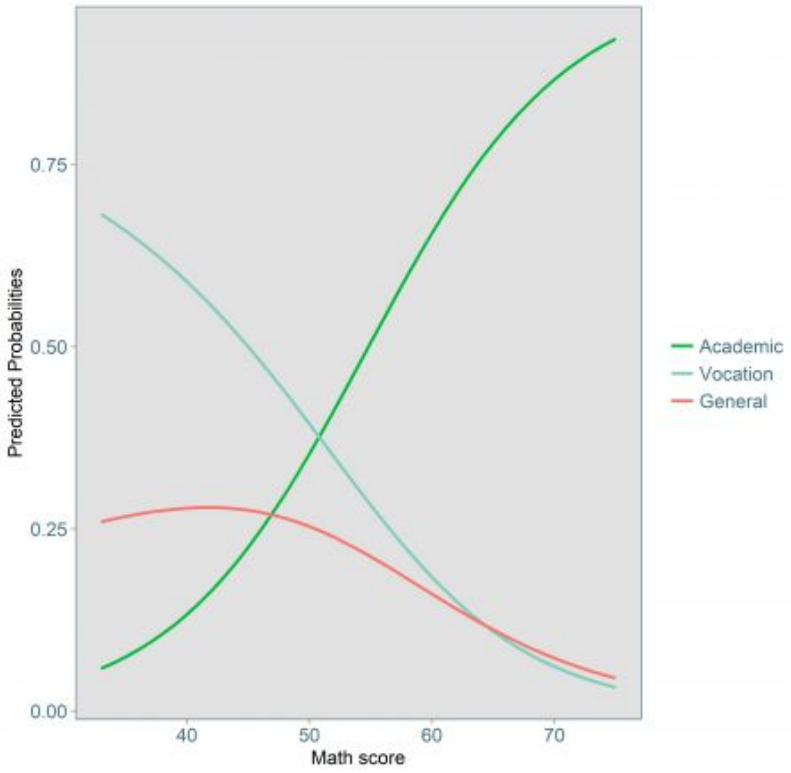
In our case, we have multiple classes (8 to be exact) and the graph looks similar to this:

Important to note that the sum of the probabilities of a single point is 1.

For example, if we look at grade 48, we can see that the probabilities are:

0.5 - Academic
0.25 - Vocation
0.25 - General

And their sum is 1



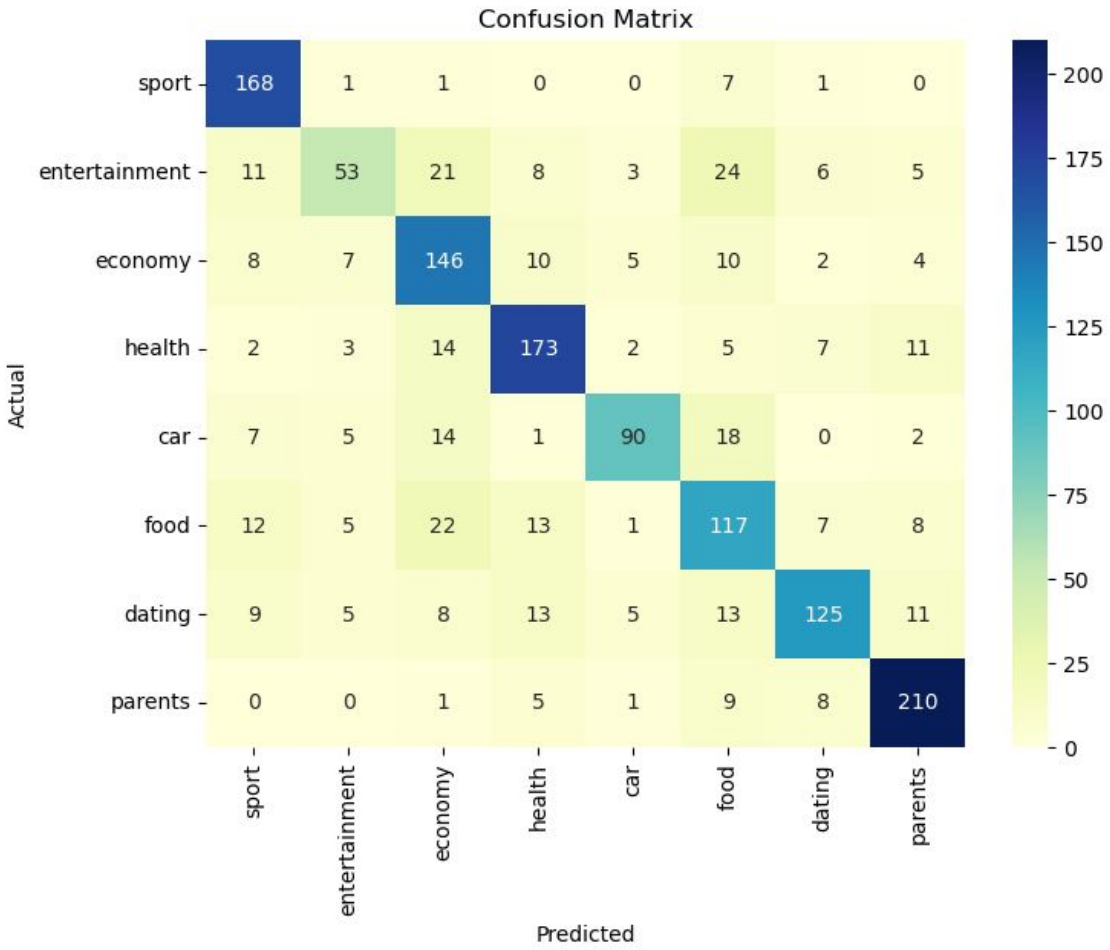
Classification Model: LogisticRegression

Our train and test process:

- Vectorize the "text" column into vectors (this is the input data X)
- Split the X and y data into X_train, X_test, y_train, y_test using train_test_split function, where the test_size is 20%
- Fit the model
- Save the model and the vectorizer into .pkl files

Confusion Matrix

Testing the model on %20 of the data and see how many times the model answered correctly



Model Performance Metrics

LogisticRegression

	precision	recall	f1-score	support
0	0.77	0.94	0.85	178
1	0.67	0.40	0.50	131
2	0.64	0.76	0.70	192
3	0.78	0.80	0.79	217
4	0.84	0.66	0.74	137
5	0.58	0.63	0.60	185
6	0.80	0.66	0.72	189
7	0.84	0.90	0.87	234
accuracy			0.74	1463
macro avg	0.74	0.72	0.72	1463
weighted avg	0.74	0.74	0.73	1463

Model Performance Metrics

Here are a few metrics that helps check model performance.

For example, let's look at the Sports category:

Precision = 168 / (168 + 49) = **0.77**

Recall = 168 / (168 + 10) = **0.94**

F1 = (2 * 0.77 * 0.94) / (0.77 + 0.94) = **0.85**

Accuracy for all = 1082 / 1463 = **0.74**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Backend + Frontend

Backend: Flask server loads the model the vectorizer and listens for requests.
Frontend: React framework, we send requests to the backend and show the results in a table.

Headers Payload Preview Response Initiator Timing


```
{0: {confidence_percentage: 0.8927070539192865, english: "Sport", hebrew: "ספורט"},…}
  0: {confidence_percentage: 0.8927070539192865, english: "Sport", hebrew: "ספורט"}
  1: {confidence_percentage: 0.016499586357066467, english: "Entertainment", hebrew: "תרבות"}
  2: {confidence_percentage: 0.017083240317317585, english: "Economy", hebrew: "כלכלה"}
  3: {confidence_percentage: 0.01433074011155041, english: "Health", hebrew: "בריאות"}
  4: {confidence_percentage: 0.01738512554484206, english: "Car", hebrew: "רכב"}
  5: {confidence_percentage: 0.020247947447207933, english: "Food", hebrew: "אוכל"}
  6: {confidence_percentage: 0.013065666603139063, english: "Dating", hebrew: "יחסים"}
  7: {confidence_percentage: 0.008680639699589945, english: "Parents", hebrew: "הורים"}
```

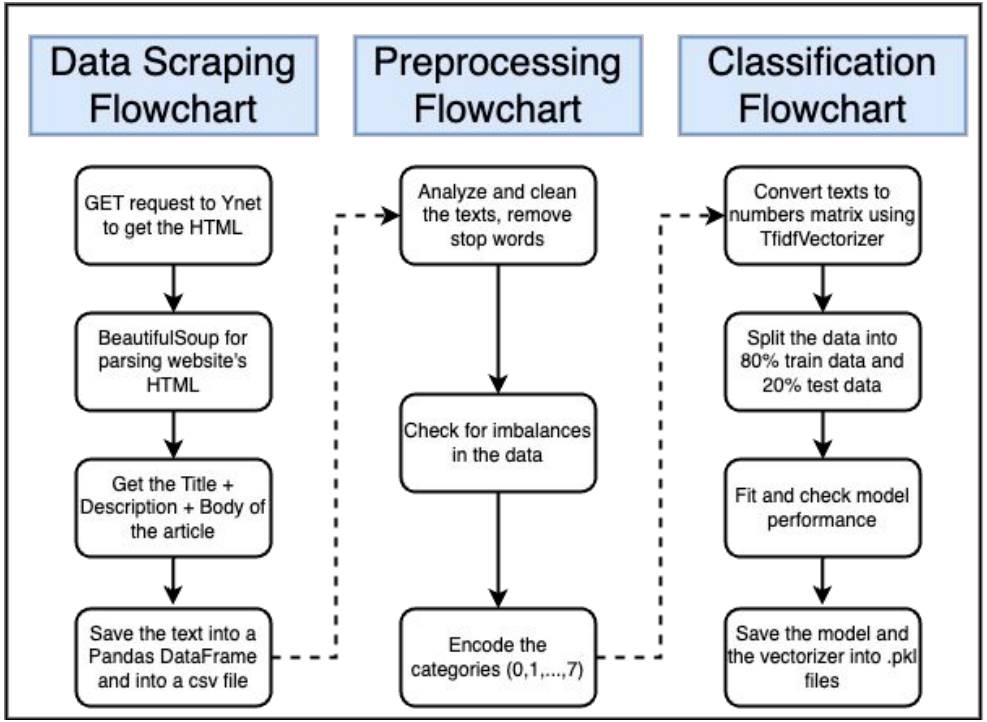

סיווג טקסטים לפי קטגוריות

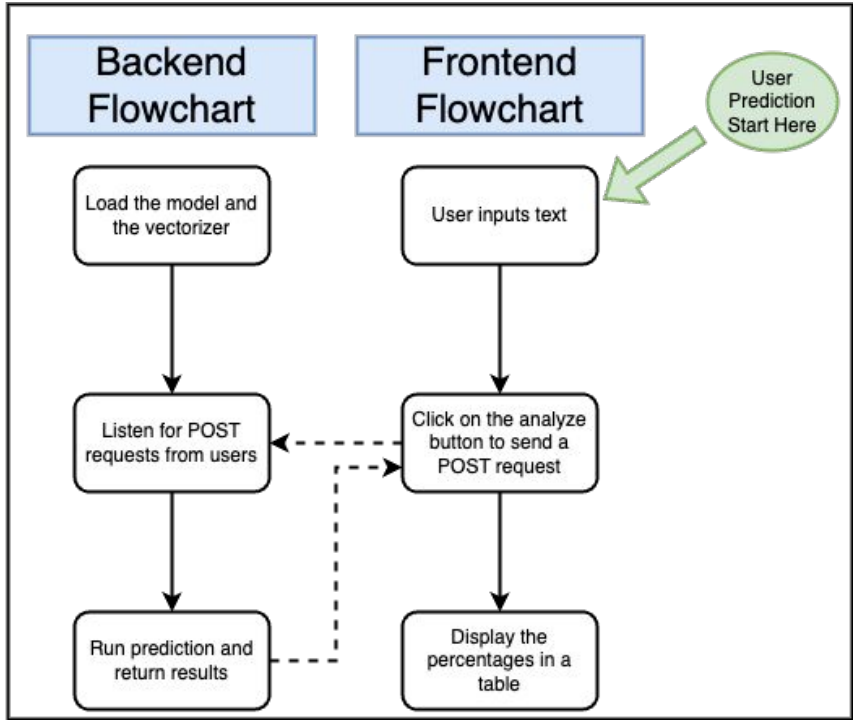
Category - קטגוריה	אחוז ודאות
Sport - ספורט	89.271 %
Food - אוכל	2.025 %
Car - רכב	1.739 %
Economy - כלכלה	1.708 %
Entertainment - תרבות	1.65 %
Health - בריאות	1.433 %
Dating - יחסים	1.307 %
Parents - הורים	0.868 %

למורסיה: מרקו תֻדוֹרוֹבִיץ' 23 נקודות, דסטין סלִיבָה 13 נק', דִילַאן אניס 10 נק', ג'ונה ראדנבו 9 נק', מוסה דיאן ורוֹדוֹנִיוֹס קוֹרוֹקוֹס 2 נק' כ"א וטרוי קוֹפִיין נקודה אחת.רבע ראשון - 16:17
לַמוֹרסִיהַחֲמִישִׁיִּת הפועל חולון: ניב משגב, דרו קרוֹפורד, סי ג'יי האריס, קווין הרוּיִ וג'סטין סמית'.חמישיית מורסיה: טרוי קוֹפִיין, דִילַאן אניס, רוֹדִיוֹנוֹס קוֹרוֹקוֹס, דסטין סלִיבָה, וּמַרקוּ תֻדוֹרוֹבִיץ'.סליבה וקוֹרוֹקוֹס היו אחראים לַחֲמֵשׁ הנקודות הראשונות של הספרדים, כְּשֶׁהרוּיִ היה הראשון לעלות על הלוח בצד של חולון. שלשה של סי ג'יי האריס השוותה את התוצאה, אַך אַנִיס הֶחזִיר מיד עם אחת משלו. ניב משגב והרוּיִ קלעו מחצי מרחק ושלשה של האריס העלה את החבורה של עמית שרף ליתרון ראשון בהתמודדות.היתרון החל להתנדנד מצד לצד כשג'ונה ראדבו קלע מעבר לקשת מהפינה ובצד השני טאג'יר מקול דייק פעמיים מהקו. הגארד של חולון נחסם במהלך הראשון של הרבע וְלוֹחַ התוצאות הראה 16:17 למורסיה לאחר 10 דקות של כדורסל.רבע שני - 31:33
לַמוֹרסִיהַשֵׁחַר עמיר קלע שלשה במהלך הראשון של הרבע, מקול קלע גם כן בפתיחה כשבצד השני ארבע נקודות של רַאדבו קבעו 21:21. האנס וַאגוּוין הוֹסיף שלשה ושתי נקודות של הרוּיִ מהקו העלו את הסגולים ליתרון חמש. ההייליט הגדול של הערב הגיע מניב משגב שהרים האלי הופ מחצי מגרש לג'סטין סמית' שסיים בְּדַאנְק. סיטו אוֹלוֹנסוֹ עצר את הדהירה של חולון עם פסק זמן.סמית' עם סל ועבירה העלה את ההפרש לְשָׁמוֹנֶה, תֻדוֹרוֹבִיץ' קלע וְהוֹסִיף נקודה מֵהקַו ונקודות של אניס צימקו את ההפרש לשלוש בלבד. תֻדוֹרוֹבִיץ' וסלִיבָה השלימו ריצת 0:10 שהחזירה את היתרון לספרדים וגרמה לעמית שרף להזעיק פסק זמן. סמית' ומקול החטיאו זריקות פנויות מתחת הסל לקראת סיום הרבע ובסיומו הישראלים ירדו בפיגור 33:31 לאחר שלא קלעו נקודות בכמעט ארבע דקותיו האחרונות.רבע שלישי - 43:56 לַמוֹרסִיהַלַּמְרוֹת נקודות של ניב משגב, פתיחת הרבע הייתה של הספרדים וְשַׁלְשָׁה של תֻדוֹרוֹבִיץ' העלתה את היתרון שלהם לַחֲמֵשׁ הפרש. הסגולים ניצלו בזכות כמה החטאות קלות

פענח

Flowchart


Data Scraping Flowchart
GET request to Ynet to get the HTML
BeautifulSoup for parsing website's HTML
Get the Title + Description + Body of the article
Save the text into a Pandas DataFrame and into a csv file
Preprocessing Flowchart
Analyze and clean the texts, remove stop words
Check for imbalances in the data
Encode the categories (0,1,...,7)
Classification Flowchart
Convert texts to numbers matrix using TfidfVectorizer
Split the data into 80% train data and 20% test data
Fit and check model performance
Save the model and the vectorizer into .pkl files


Backend Flowchart
Load the model and the vectorizer
Listen for POST requests from users
Run prediction and return results
Frontend Flowchart
User Prediction Start Here
User inputs text
Click on the analyze button to send a POST request
Display the percentages in a table

Technologies

- Data scraping: BeautifulSoup, Pandas
- Preprocessing: NLP, Imbalanced data, Encoding
- Classification model: TfidfVectorizer, LogisticRegression
- Backend: Flask
- Frontend: React

Sport - Google Docs | Food - Google Docs | Dating - Google Docs | Vite + React

127.0.0.1:5173

סיווג טקסטים לפי קטגוריות

Category - קטגוריה	אחוז ודאות ↓
Sport - ספורט	0 %
Entertainment - תרבות	0 %
Economy - כלכלה	0 %
Health - בריאות	0 %
Car - רכב	0 %
Food - אוכל	0 %
Dating - יחסים	0 %
Parents - הורים	0 %

יש להזין את הטקסט...

פענח